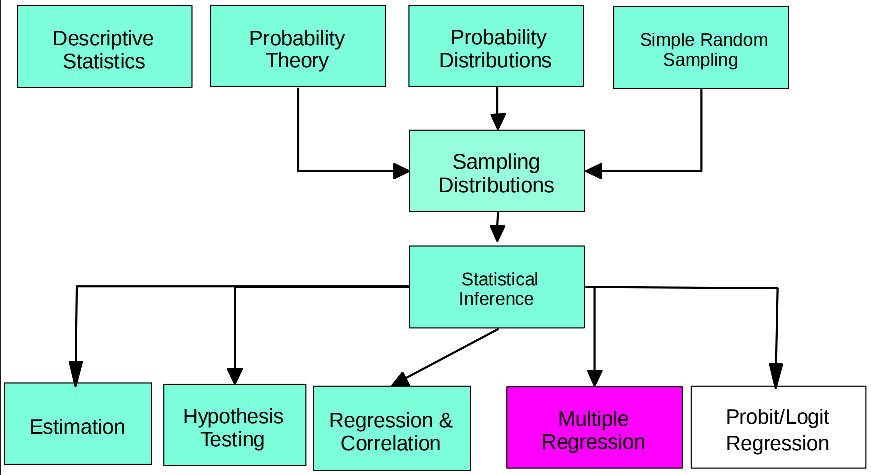
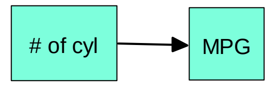
The Course So Far:



- ✓ Number of cylinders
- ✓ Engine displacement
- ✓ Horsepower
- ✓ Acceleration
- ✓ Price (?)

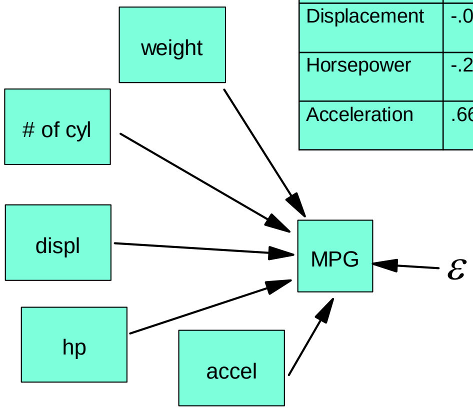



```
.  regress  mpg cylinder

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  126.87
       Model |  3788.24488     1  3788.24488           Prob > F      =  0.0000
    Residual |  4538.50863   152  29.8586094           R-squared     =  0.4549
-------------+------------------------------           Adj R-squared =  0.4514
       Total |  8326.75351   153  54.4232255           Root MSE      =  5.4643

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cylinder |  -3.637801   .3229644   -11.26   0.000    -4.275879   -2.999722
       _cons |   46.41558   1.625274    28.56   0.000     43.20454    49.62662
------------------------------------------------------------------------------
```

```
.  regress  mpg  displace

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =  201.22
       Model |  4743.50777     1  4743.50777           Prob > F      =  0.0000
    Residual |  3583.24573   152  23.5739851           R-squared     =  0.5697
-------------+------------------------------           Adj R-squared =  0.5668
       Total |  8326.75351   153  54.4232255           Root MSE      =  4.8553

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    displace |  -.0765923   .0053995   -14.19   0.000    -.0872601   -.0659246
       _cons |   40.56734   .9176048    44.21   0.000     38.75444    42.38025
------------------------------------------------------------------------------


.  regress  mpg   horsepow

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  1,   148) =  243.66
       Model |  5030.94725     1  5030.94725           Prob > F      =  0.0000
    Residual |  3055.82715   148  20.6474807           R-squared     =  0.6221
-------------+------------------------------           Adj R-squared =  0.6196
       Total |   8086.7744   149  54.2736537           Root MSE      =  4.5439

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    horsepow |   -.237707   .0152283   -15.61   0.000    -.2677999    -.207614
       _cons |   49.87064   1.403124    35.54   0.000      47.0979    52.64339
------------------------------------------------------------------------------
```

```
. regress  mpg accel

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =    8.26
       Model |  429.143083     1  429.143083           Prob > F      =  0.0046
    Residual |  7897.61042   152  51.9579633           R-squared     =  0.0515
-------------+------------------------------           Adj R-squared =  0.0453
       Total |  8326.75351   153  54.4232255           Root MSE      =  7.2082

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       accel |   .6626446   .2305715     2.87   0.005     .2071058    1.118183
       _cons |   18.00445   3.798794     4.74   0.000     10.49919     25.5097
------------------------------------------------------------------------------

. regress  mpg price

      Source |       SS       df       MS              Number of obs =     154
-------------+------------------------------           F(  1,   152) =    0.00
       Model |  .007485901     1  .007485901           Prob > F      =  0.9907
    Residual |  8326.74602   152  54.7812238           R-squared     =  0.0000
-------------+------------------------------           Adj R-squared = -0.0066
       Total |  8326.75351   153  54.4232255           Root MSE      =  7.4014

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       price |  -3.47e-06   .0002971    -0.01   0.991    -.0005905    .0005836
       _cons |   28.80952    1.49389    19.28   0.000     25.85805    31.76099
------------------------------------------------------------------------------
```

Regression	Coefficient (t-statistic), R-squared
Weight	-.0101428 (-18.28) R-square = 0.6874
Cylinders	-3.637801 (-11.26) R-square = 0.4549
Displacement	-.0765923 (-14.19) R-square = 0.5697
Horsepower	-.237707 (-15.61) R-square = 0.6221
Acceleration	.6626446 (2.87) R-square = 0.0515



$$MPG = \beta_0 + \beta_1 \bullet wgt + \beta_2 \bullet cyl + \beta_3 \bullet displ + \beta_4 \bullet hp + \beta_5 \bullet accel + \varepsilon$$

```
. regress mpg  weight cylinder displace horsepow accel

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  5,   144) =   79.87
       Model |  5943.51201     5  1188.7024            Prob > F      =  0.0000
    Residual |  2143.26239   144  14.8837666           R-squared     =  0.7350
-------------+------------------------------           Adj R-squared =  0.7258
       Total |   8086.7744   149  54.2736537           Root MSE      =  3.8579

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0101399   .0017803    -5.70   0.000    -.0136589   -.0066209
    cylinder |   .2859285   .6469444     0.44   0.659    -.9928057    1.564663
    displace |   .0182705   .0173536     1.05   0.294    -.0160301    .0525712
    horsepow |  -.0733252   .0355269    -2.06   0.041    -.1435468   -.0031036
       accel |   .3602982   .2056064     1.75   0.082    -.0460983    .7666947
       _cons |   52.32295   3.970578    13.18   0.000      44.4748    60.17109
------------------------------------------------------------------------------


. correlate  cylinder displace horsepow accel weight
(obs=150)

             | cylinder displace horsepow    accel   weight
-------------+---------------------------------------------
    cylinder |   1.0000
    displace |   0.9318   1.0000
    horsepow |   0.7779   0.8226   1.0000
       accel |  -0.2101  -0.2066  -0.4929   1.0000
      weight |   0.8300   0.9150   0.8107  -0.0478   1.0000
```

```
. regress  mpg weight cylinder displace horsepow accel , beta

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  5,   144) =   79.87
       Model |  5943.51201     5  1188.7024            Prob > F      =  0.0000
    Residual |  2143.26239   144  14.8837666           R-squared     =  0.7350
-------------+------------------------------           Adj R-squared =  0.7258
       Total |   8086.7744   149  54.2736537           Root MSE      =  3.8579

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      weight |  -.0101399   .0017803    -5.70   0.000                -.8333565
    cylinder |   .2859285   .6469444     0.44   0.659                 .0535203
    displace |   .0182705   .0173536     1.05   0.294                 .1817973
    horsepow |  -.0733252   .0355269    -2.06   0.041                -.2433036
       accel |   .3602982   .2056064     1.75   0.082                 .1237697
       _cons |   52.32295   3.970578    13.18   0.000                        .
------------------------------------------------------------------------------
```

Standardized regression coefficients show *relative* importance of each independent variable

```
. regress  mpg weight cylinder displace horsepow accel , beta (Turchi)

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  5,   144) =   79.87
       Model |  5943.51201     5  1188.7024            Prob > F      =  0.0000
    Residual |  2143.26239   144  14.8837666           R-squared     =  0.7350
-------------+------------------------------           Adj R-squared =  0.7258
       Total |   8086.7744   149  54.2736537           Root MSE      =  3.8579

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
      weight |  -.0101399   .0017803    -5.70   0.000                -.8333565
    cylinder |   .2859285   .6469444     0.44   0.659                 .0535203
    displace |   .0182705   .0173536     1.05   0.294                 .1817973
    horsepow |  -.0733252   .0355269    -2.06   0.041                -.2433036
       accel |   .3602982   .2056064     1.75   0.082                 .1237697
       _cons |   52.32295   3.970578    13.18   0.000                        .
------------------------------------------------------------------------------

. regress  mpg cylinder displace horsepow accel year weight price, beta (Gasana)

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  7,   142) =   68.90
       Model |  6247.30769     7  892.472528           Prob > F      =  0.0000
    Residual |  1839.46671   142  12.9539909           R-squared     =  0.7725
-------------+------------------------------           Adj R-squared =  0.7613
       Total |   8086.7744   149  54.2736537           Root MSE      =  3.5992

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|                     Beta
-------------+----------------------------------------------------------------
    cylinder |   .4368451   .6167177     0.71   0.480                  .081769
    displace |   .0287842   .0170537     1.69   0.094                 .2864113
    horsepow |  -.0554725   .0333758    -1.66   0.099                -.1840657
       accel |    .431885    .193399     2.23   0.027                 .1483612
        year |   .2809669   .2709474     1.04   0.302                   .05588
      weight |  -.0123731   .0018463    -6.70   0.000                -1.016897
       price |   .0006226    .000195     3.19   0.002                 .1718182
       _cons |   27.85781    22.5288     1.24   0.218                        .
------------------------------------------------------------------------------
```

1. Inclusion of price makes acceleration significant

2. R-square rises

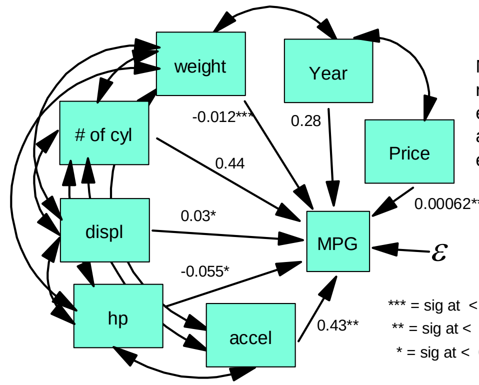
3. Engine displacement's relative effect rises

4. Price is now significant unlike the bivariate regression!

Handout: Multiple Regression Analysis of Mileage Data

```
. regress  mpg cylinder displace horsepow accel year weight price, beta
(Gasana)

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  7,   142) =   68.90
       Model |  6247.30769     7  892.472528           Prob > F      =  0.0000
    Residual |  1839.46671   142  12.9539909           R-squared     =  0.7725
-------------+------------------------------           Adj R-squared =  0.7613
       Total |   8086.7744   149  54.2736537           Root MSE      =  3.5992

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.       t     P>|t|                  Beta
-------------+----------------------------------------------------------------
    cylinder |   .4368451   .6167177     0.71   0.480                 .081769
    displace |   .0287842   .0170537     1.69   0.094                .2864113
    horsepow |  -.0554725   .0333758    -1.66   0.099               -.1840657
       accel |    .431885    .193399     2.23   0.027                .1483612
        year |   .2809669   .2709474     1.04   0.302                  .05588
      weight |  -.0123731   .0018463    -6.70   0.000               -1.016897
       price |   .0006226    .000195     3.19   0.002                .1718182
       _cons |   27.85781    22.5288     1.24   0.218                       .
------------------------------------------------------------------------------
```

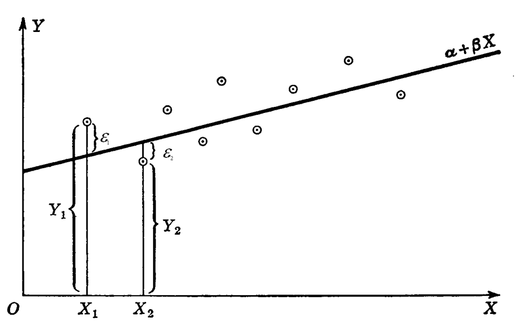
Regression	Simple Regression Coefficient (all significant except for price)	Multiple Regression Coefficient
Weight	-.010	-0.012***
Cylinders	-3.63	0.44
Displacement	-.076	0.03*
Horsepower	-.23	-0.055*
Acceleration	.66	0.43**
Year	2.122	0.28
Price	-3.47e-06	0.00062***



Multiple regression is required when relevant explanatory variables are correlated with each other.

*** = sig at < 0.01 level
** = sig at < 0.05 level
* = sig at < 0.10 level

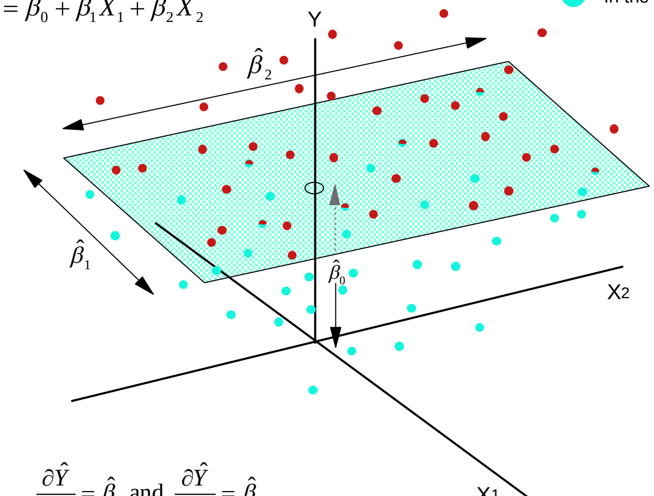
General Rule: Use multiple regression when a number of variables have an effect on a dependent variable *and* those variables are correlated with each other, such that leaving one of them out of the regression leads to the included variables' coefficients reflecting some of the excluded variable's influences.



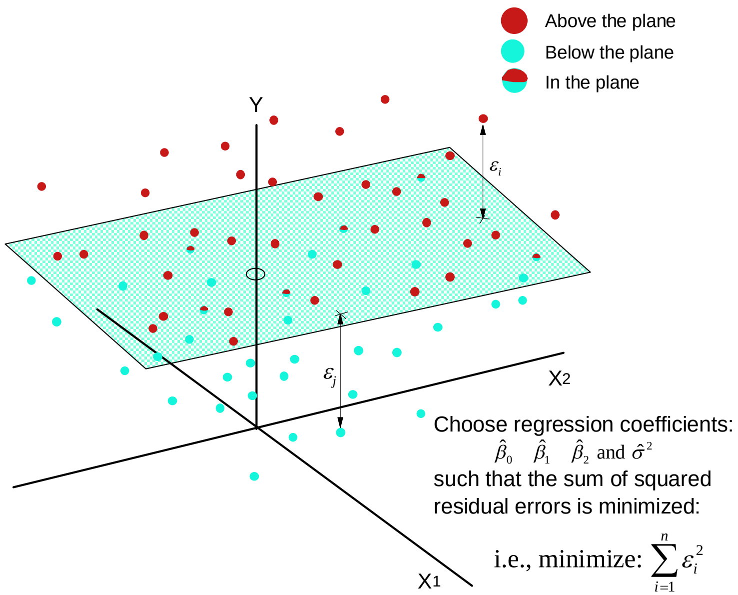
Fitting a multiple regression plane through a 3-dimensional scatter plot

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2$$

Above the plane
Below the plane
In the plane



$$\frac{\partial \hat{Y}}{\partial X_1} = \hat{\beta}_1 \text{ and } \frac{\partial \hat{Y}}{\partial X_2} = \hat{\beta}_2$$



Choose regression coefficients:

$\hat{\beta}_0 \quad \hat{\beta}_1 \quad \hat{\beta}_2$ and $\hat{\sigma}^2$

such that the sum of squared residual errors is minimized:

i.e., minimize: $\sum_{i=1}^{n} \varepsilon_i^2$

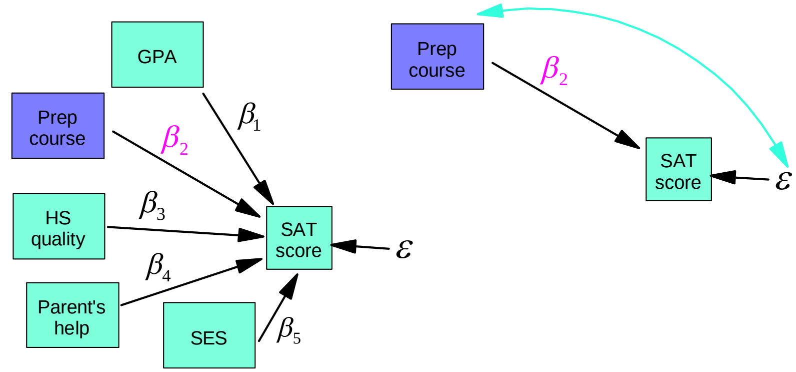
✓ Ability as measured, say, by high school GPA

✓ Test-taking skill as augmented by having taken an SAT prep. course.

✓ Quality of high school education as measured by some sort of a school effectiveness score.

✓ Parental assistance in preparing for test.

✓ Socioeconomic level (SES) of student as measured, say, by parents' income and education.



✓ Find a [very] large number of students and assign them into homogeneous groups; that is, into groups where every student is the same as every other student with respect to GPA, high school quality, parental help, SES.

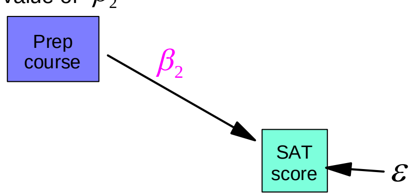
✓ Then, in each group *randomly* assign one half the students to take the prep course and leave the other group without the prep course as a ***control group.***

✓ Test the students before and after the course to measure gains (remember the control group will, on average, see a gain in scores even without the prep course).

✓ Do a simple regression to determine the value of β_2

By controlling the values of all other variables, we remove the correlation between the error term and *prep course* and the beta coefficient becomes un-biased:

$$E\left[\hat{\beta}_2\right] = \beta_2$$



Problems with social experiments:

✓ *Failure to randomize.* It's often difficult, or even impossible, to randomly assign observations into two groups. Difficult because our data are not often experimental, but are essentially made available to us without any possibility of arranging them into any sort of *experimental design.*

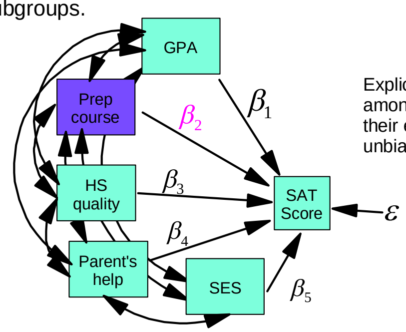
✓ *Failure to follow the "treatment" protocol.* Even if we can randomly assign students to the prep course -- or not -- we cannot make the students study or otherwise follow the training protocol. Also, we cannot stop control group subjects from taking the course on the sly.

✓ *Attrition.* Subjects drop out of experiments, and their tendency to drop out is not random. That is, non-random attrition may lead to correlation between errors and treatment variable (known as **selection bias**).

✓ *Cost and small samples.* Because experiments with humans are very expensive, the samples tend to be small, reducing the precision of our statistical estimation.

Advantages of Multiple Regression:

- ✓ By explicitly including the other, relevant and correlated, explanatory variables, it enables us to control for them without requiring them to be exactly the same for any two people. That is, by explicitly taking account of the intercorrelations among the potential explanatory variables, multiple regression allow us *to control for their influence without requiring them to take on only a restricted set of values.*
- ✓ It allows for the simultaneous control of many variables even though no two people are exactly alike on all the variables.
- ✓ It allows us to generate a single estimate for the "effect" of each explanatory variable, which is analogous to the weighted average of effects in different subgroups.



Explicitly taking account of the intercorrelations among the explanatory variables serves to "control" their effects, so that β_2 represents the "true" (or unbiased) effect of the prep course on SAT score.

Potential Problems in using Multiple Regression to "Control" other variables:

- ✓ We have to assume an explicit functional relationship among the variables -- it could be wrong.
- ✓ Unlike a randomized experiment, we have to be able to *measure* the explanatory variable to be able to include it in the statistical analysis. In addition, the variable must be measured *well* to avoid "measurement error."
- ✓ We have to include all the relevant and correlated explanatory variables in order to avoid **omitted variable bias.** Randomization controls for all characteristics of the experimental subjects, regardless of whether those characteristics can be measured.

Definition: A statistical analysis is **internally valid** if the statistical inferences about causal effects are valid for the population being studied. The analysis is **externally valid** if its inferences and conclusions can be generalized from the population and setting studied to other populations and settings.

Internal validity has two components:
The estimator of the causal effect should be unbiased and consistent. That is, a slope coefficient $\hat{\beta}_i$ should be an unbiased and consistent estimator of the true population effect, β. That is, $E\left[\hat{\beta}_i\right] = \beta$.

Hypothesis tests should have the desired significance level (the actual rejection rate of the test under the null hypothesis should equal its desired significance level).

So, for OLS regression internal validity requires that the OLS estimator is unbiased and consistent and that standard errors are computed in a way that makes confidence intervals have the desired confidence level.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

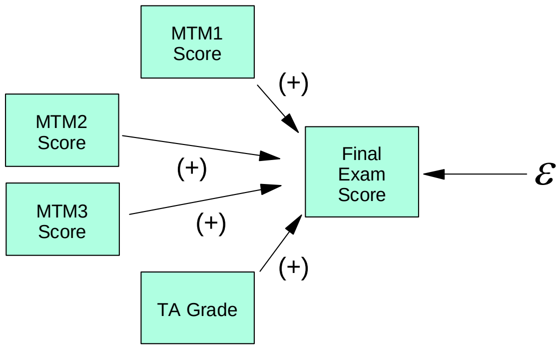
ε is a random error, that, for any given set of values of indepent variables $X_1, X_2, X_3 \ldots$ is normally distributed with mean zero and variance equal to σ_ε^2.

The random errors, ε_j and ε_k associated with any pair of *y*-values are independent of each other and independent of the values of *x*-variables included in the model.

Then, it follows that: $E\left[Y\right] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$

The model is "well specified", i.e., all relevant (and correlated) explantory variables are included in the model.

The Problem: Does performance on midterm exams predict performance on the final exam very well? Do different midterms have a different impact on final exam performance?

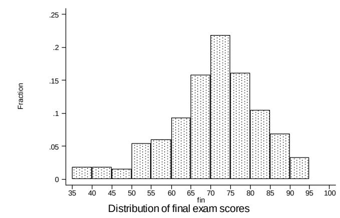


$$fin = \beta_0 + \beta_1 \bullet mtm1 + \beta_2 \bullet mtm2 + \beta_3 \bullet mtm3 + \beta_4 \bullet tag + \varepsilon$$

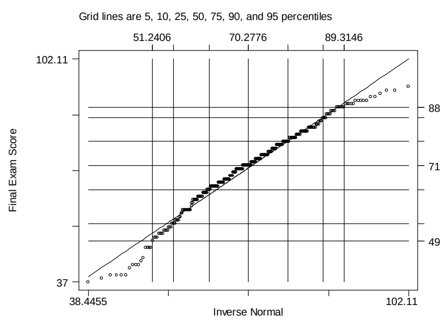
```
. describe

Contains data from D:\Amifiles\Econ 70\lectures_new\MultReg_Ec10.dta
  obs:           344                              Economics 10 Data
 vars:             6                              17 Apr 2003 17:10
 size:        16,512 (98.4% of memory free)
-------------------------------------------------------------------------------
              storage  display     value
variable name   type   format      label      variable label
-------------------------------------------------------------------------------
mtm1            double %12.0g                 First Midterm Score
mtm2            double %12.0g                 Second Midterm Score
mtm3            double %12.0g                 Third Midterm Score
tag             double %12.0g                 TA grade
fin             double %12.0g                 Final Exam Score
id              float  %9.0g                  Student ID
-------------------------------------------------------------------------------
Sorted by:  mtm1
```

```
. summarize  fin, detail

                             fin
-------------------------------------------------------------
      Percentiles      Smallest
 1%           39             37
 5%           49             38
10%           54             39       Obs                 335
25%           64             39       Sum of Wgt.         335

50%           71                      Mean           70.27761
                        Largest       Std. Dev.      11.57366
75%           78             92
90%           85             93       Variance       133.9496
95%           88             93       Skewness      -.5278463
99%           92             94       Kurtosis       3.173458
```

graph fin, bin(13) xlabel(35 40 to 100) ylabel(0 .05 to .25) ti("Distribution of final exam scores")

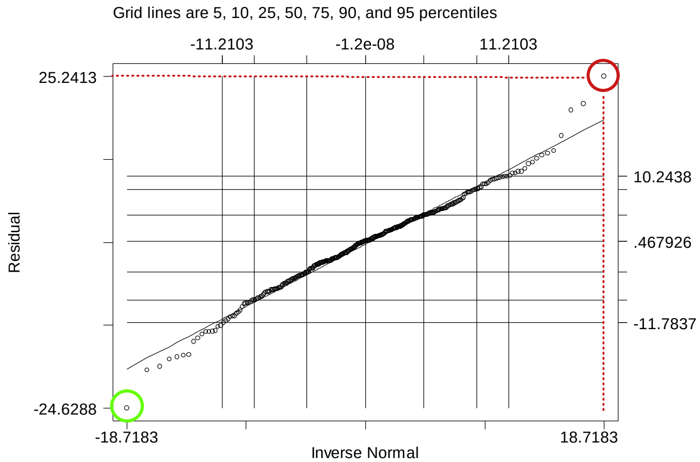


qnorm fin, grid

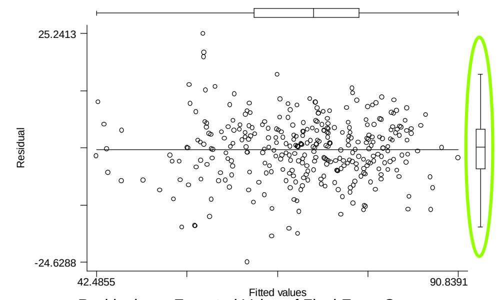



```
regress  fin mtm1 mtm2 mtm3 tag

      Source |       SS       df       MS              Number of obs =     331
-------------+------------------------------           F(  4,   326) =  155.11
       Model |  29172.9861     4  7293.24652           Prob > F      =  0.0000
    Residual |  15328.4037   326  47.0196431           R-squared     =  0.6556
-------------+------------------------------           Adj R-squared =  0.6513
       Total |  44501.3897   330  134.852696           Root MSE      =  6.8571

------------------------------------------------------------------------------
         fin |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        mtm1 |    .227831   .0463552     4.91   0.000     .1366379    .3190241
        mtm2 |   .2044693   .0395619     5.17   0.000     .1266404    .2822982
        mtm3 |   .4081122   .0370606    11.01   0.000      .335204    .4810203
         tag |    .016204    .056534     0.29   0.775    -.0950135    .1274215
       _cons |   7.930885   5.164574     1.54   0.126    -2.229214    18.09098
------------------------------------------------------------------------------

.
```

```
qnorm  epsilon, grid
```

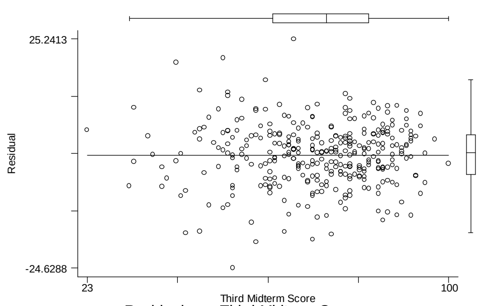



```
gr epsilon finhat, two box yline(0)
ti("Residuals vs. Expected Value of
Final Exam Score")
```

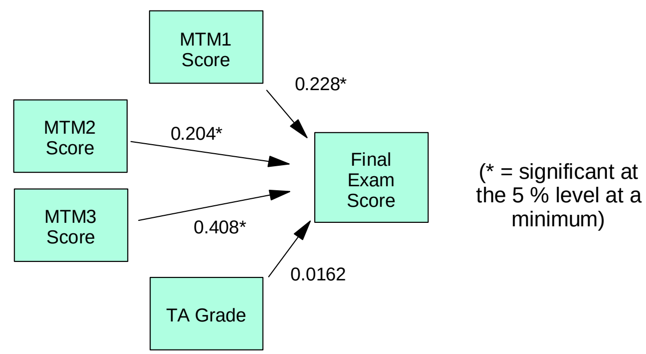


Residuals vs. Expected Value of Final Exam Score

```
gr epsilon  mtm3, two box
yline(0) ti("Residuals vs. Third
Midterm Score")
```



Residuals vs. Third Midterm Score

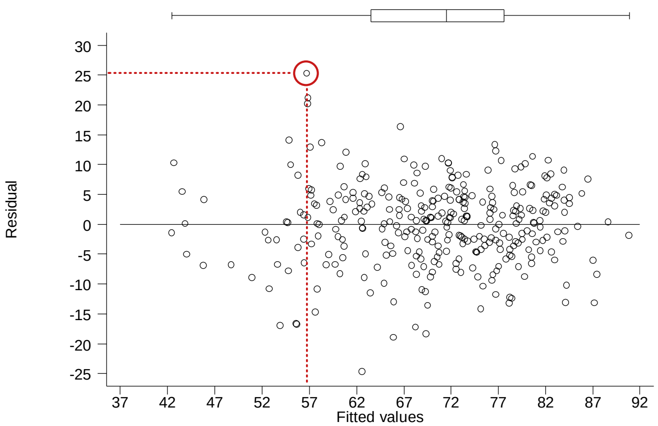


$$\widehat{fin} = \hat{\beta}_0 + \hat{\beta}_1 \bullet mtm1 + \hat{\beta}_2 \bullet mtm2 + \hat{\beta}_3 \bullet mtm3 + \hat{\beta}_4 \bullet tag$$

$$\widehat{fin} = \underset{(1.54)}{7.93} + \underset{(4.91)}{0.228} \bullet mtm1 + \underset{(5.17)}{0.204} \bullet mtm2 + \underset{(11.01)}{0.408} \bullet mtm3 + \underset{(0.29)}{0.0162} \bullet tag$$

$$R^2 = 0.6556$$

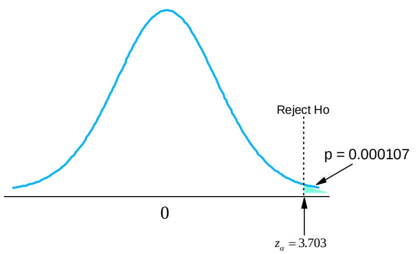
```
gr epsilon finhat, two box yline(0)
ti("Residuals vs. Expected Value of
Final Exam Score") xlabel(37 42 to 92)
ylabel(-25 -20 to 30)
```



Residuals vs. Expected Value of Final Exam Score

```
Stem-and-leaf plot for fin (Final Exam Score)

  3s | 7
  3. | 89999
  4* | 1
  4t | 2223
  4f | 4
  4s | 7777
  4. | 9
  5* | 0001111
  5t | 2222333
  5f | 44445555
  5s | 6677
  5. | 888888888889
  6* | 001111111
  6t | 2222222223333333
  6f | 44444455555555555555
  6s | 66666666666677777777777
  6. | 88888889999999
  7* | 000000000000000011111111111111111
  7t | 2222222222333333333333
  7f | 44444444444444455555555
  7s | 666666666677777777777
  7. | 8888888888888999999999999
  8* | 000000011111111111
  8t | 222222222223333
  8f | 4445555
  8s | 66667777
  8. | 88888899999
  9* | 0000011
  9t | 233
  9f | 4
```

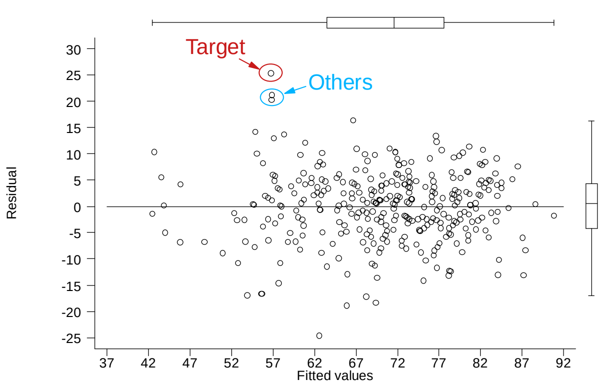
$$z = \frac{x-\mu}{\sigma} = \frac{25.241-0}{6.815} = 3.704$$



The probability of getting a residual this large is approximately one in ten thousand.

Possibilities:

✓He made a miraculous improvement on the final
✓He cheated by bringing crib sheets into the final exam
✓He cheated by having someone take the test for him
✓He cheated by copying another student's answers



Residuals vs. Expected Value of Final Exam Score

Types of variables used in regression analysis:

1. A **metric (or quantitative) variable** is a variable that is measured on some well-defined scale. For variables measured on a particular scale, a change of a certain amount means the same thing no matter where one starts. (e.g., growing one inch in height means the same thing whether or not one starts at 5'2" or 6'2")

1 = strongly disagree
2 = disagree
3 = agree
4 = strongly agree

2. This variable is an **ordinal variable.** Its values ascend or descend, but the distance between them is arbitrary and undefined. That is, 1<2<3<4, but we don't know by how much.

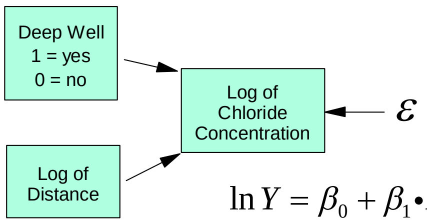
3. **Categorical variables** are variables that do not have any order at all; that is, they are measured on **nominal scales** where each value represents a different category, but the categories themselves cannot be ranked. e.g., 1 = Male and 0 = Female for the variable "sex."

A **dummy variable** is a categorical variable that can take on only one of two values, zero or one.

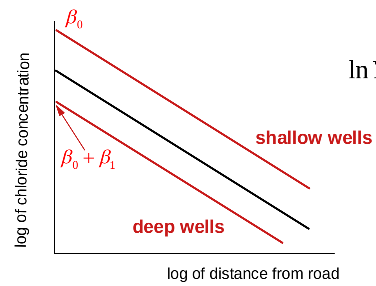
Study of Road Salt Contamination of New England Wells

Variables

1. *lnclor:* natural logarithm of the chloride concentration (log milligrams per liter) in the well's water -- an indicator of contamination by road salt.
2. *deep:* dummy variable coded 0 for shallow wells and 1 for deeper wells drilled into bedrock.
3. *lndroad:* natural logarithm of the distance (log feet) between the well and the nearest salted road.



$$\ln Y = \beta_0 + \beta_1 \bullet Deep + \beta_2 \bullet \ln droad + \varepsilon$$



$$\ln Y = \beta_0 + \beta_1 \bullet Deep + \beta_2 \bullet \ln droad + \varepsilon$$

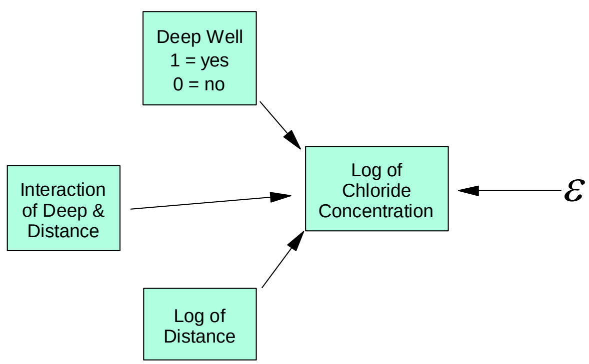
The dummy variable *Deep* is an intercept dummy variable; it serves to adjust the intercept to account for the effect of the depth of the wells, but it does not affect the impact of distance from the road.

```
. regress  lnchlor deep lndroad

      Source |       SS       df       MS              Number of obs =      52
-------------+------------------------------           F(  2,    49) =    1.21
       Model |  4.50187596     2  2.25093798           Prob > F      =  0.3084
    Residual |  91.5203226    49  1.86776169           R-squared     =  0.0469
-------------+------------------------------           Adj R-squared =  0.0080
       Total |  96.0221986    51  1.88278821           Root MSE      =  1.3667

------------------------------------------------------------------------------
     lnchlor |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        deep |  -.6971194   .4811856    -1.45   0.154    -1.664098    .2698592
     lndroad |  -.0909673   .1797176    -0.51   0.615    -.4521233    .2701886
       _cons |    4.20954   .9609568     4.38   0.000     2.278425    6.140655
------------------------------------------------------------------------------
```

```
generate deeproad = deep*lndroad
```

$$\ln Y = \beta_0 + \beta_1 \bullet Deep + \beta_2 \bullet \ln droad + \varepsilon$$

$$\ln Y = \beta_0 + \beta_1 \bullet Deep + \beta_2 \bullet \ln droad + \beta_3 \bullet deeproad + \varepsilon$$



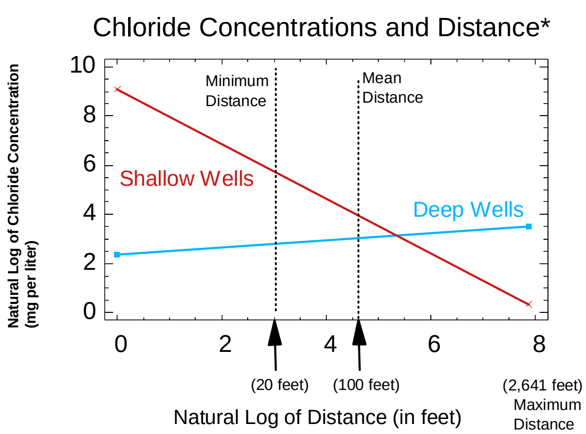
$$\ln Y = \beta_0 + \beta_1 \bullet Deep + \beta_2 \bullet \ln droad + \beta_3 \bullet deep \times \ln droad$$

$$\frac{\partial \ln Y}{\partial \ln droad} = \beta_2 + \beta_3 \bullet deep$$

```
. regress  lnchlor deep lndroad deeproad

      Source |       SS       df       MS              Number of obs =      52
-------------+------------------------------           F(  3,    48) =    3.81
       Model |  18.4831272     3  6.1610424            Prob > F      =  0.0157
    Residual |  77.5390714    48  1.61539732           R-squared     =  0.1925
-------------+------------------------------           Adj R-squared =  0.1420
       Total |  96.0221986    51  1.88278821           Root MSE      =   1.271

------------------------------------------------------------------------------
     lnchlor |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        deep |  -6.717366   2.094713    -3.21   0.002    -10.92907   -2.505663
     lndroad |  -1.109424   .3844204    -2.89   0.006    -1.882354   -.3364954
    deeproad |   1.255847   .4268777     2.94   0.005     .3975521    2.114143
       _cons |   9.073459   1.879384     4.83   0.000     5.294704    12.85221
------------------------------------------------------------------------------
```

Distance matters!

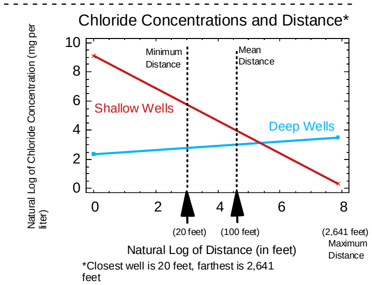
But only if the well is not deep!



*Closest well is 20 feet, farthest is 2,641 feet

```
. regress  lnchlor deep lndroad deeproad

      Source |       SS       df       MS              Number of obs =      52
-------------+------------------------------           F(  3,    48) =    3.81
       Model |  18.4831272     3  6.1610424            Prob > F      =  0.0157
    Residual |  77.5390714    48  1.61539732           R-squared     =  0.1925
-------------+------------------------------           Adj R-squared =  0.1420
       Total |  96.0221986    51  1.88278821           Root MSE      =   1.271

------------------------------------------------------------------------------
     lnchlor |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        deep |  -6.717366   2.094713    -3.21   0.002    -10.92907   -2.505663
     lndroad |  -1.109424   .3844204    -2.89   0.006    -1.882354   -.3364954
    deeproad |   1.255847   .4268777     2.94   0.005     .3975521    2.114143
       _cons |   9.073459   1.879384     4.83   0.000     5.294704    12.85221
------------------------------------------------------------------------------
```

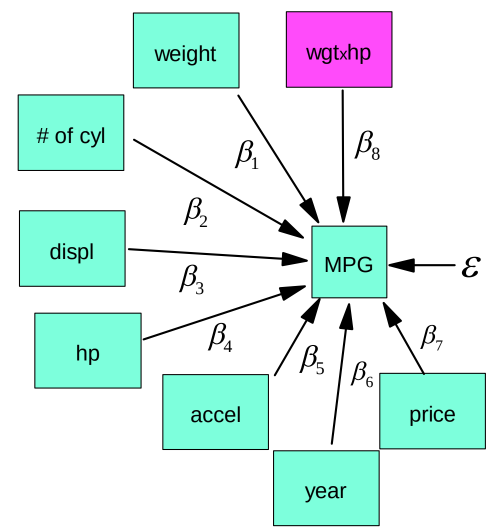
Handout on using dummy variables to test interaction effects in regression analysis




```
. regress   mpg weight cylinder displace horsepow accel year  price

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  7,   142) =   68.90
       Model |  6247.30769     7  892.472528           Prob > F      =  0.0000
    Residual |  1839.46671   142  12.9539909           R-squared     =  0.7725
-------------+------------------------------           Adj R-squared =  0.7613
       Total |   8086.7744   149  54.2736537           Root MSE      =  3.5992

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0123731   .0018463    -6.70   0.000    -.0160229   -.0087234
    cylinder |   .4368451   .6167177     0.71   0.480    -.7822891    1.655979
    displace |   .0287842   .0170537     1.69   0.094    -.0049278    .0624962
    horsepow |  -.0554725   .0333758    -1.66   0.099    -.1214501    .0105051
       accel |    .431885    .193399     2.23   0.027     .0495716    .8141984
        year |   .2809669   .2709474     1.04   0.302    -.2546449    .8165786
       price |   .0006226    .000195     3.19   0.002     .0002371    .0010081
       _cons |   27.85781    22.5288     1.24   0.218    -16.67736    72.39298
------------------------------------------------------------------------------
```

. generate wthp= weight* horsepow

$$MPG = \beta_0 + \beta_1 \cdot wgt + \beta_2 \cdot cyl + \beta_3 \cdot displ + \beta_4 \cdot hp + \beta_5 \cdot accel + \beta_6 \cdot year + \beta_7 \cdot price + \beta_8 \cdot wgt \times hp + \varepsilon$$



```
. regress   mpg weight cylinder displace horsepow accel year  price  wthp

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  8,   141) =   66.61
       Model |  6394.74812     8  799.343515           Prob > F      =  0.0000
    Residual |  1692.02628   141  12.0001864           R-squared     =  0.7908
-------------+------------------------------           Adj R-squared =  0.7789
       Total |   8086.7744   149  54.2736537           Root MSE      =  3.4641

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0167031   .0021642    -7.72   0.000    -.0209816   -.0124246
    cylinder |   .0609603    .603188     0.10   0.920    -1.131501    1.253422
    displace |   .0156661   .0168351     0.93   0.354    -.0176157     .048948
    horsepow |  -.2849999   .0729368    -3.91   0.000     -.429191   -.1408088
       accel |   .3059297   .1895796     1.61   0.109    -.0688561    .6807156
        year |   .3410433   .2613443     1.30   0.194    -.1756165    .8577031
       price |    .000649   .0001878     3.45   0.001     .0002776    .0010203
        wthp |   .0000745   .0000212     3.51   0.001     .0000325    .0001165
       _cons |   42.22624    22.0676     1.91   0.058    -1.399893    85.85238
------------------------------------------------------------------------------
```

$$\frac{\partial MPG}{\partial wgt} = \beta_1 + \beta_8 \bullet hp$$

The marginal impact of a one pound change in a car's weight depends on the horsepower of the engine in the car.

$$\frac{\partial MPG}{\partial wgt} = \beta_1 + \beta_8 \bullet hp = -0.0167 + 0.0000745 \times hp$$

$$\frac{\partial MPG}{\partial wgt} = \beta_1 + \beta_8 \bullet hp = -0.0167 + 0.0000745 \times 60 = -0.012$$

$$\frac{\partial MPG}{\partial wgt} = \beta_1 + \beta_8 \bullet hp = -0.0167 + 0.0000745 \times 139 = -0.006$$

$$\frac{\partial MPG}{\partial hp} = \beta_4 + \beta_8 \bullet wgt = -0.285 + 0.0000745 \times wgt$$

$$\frac{\partial MPG}{\partial hp} = \beta_4 + \beta_8 \bullet wgt = -0.285 + 0.0000745 \times 1{,}875 = -0.145$$

$$\frac{\partial MPG}{\partial hp} = \beta_4 + \beta_8 \bullet wgt = -0.285 + 0.0000745 \times 3{,}830 = 0.000335$$

	No Interaction		With Interaction	
	Coef.	t	Coef.	t
weight	**-.0123731**	**-6.70**	**-.0167031**	**-7.72**
cylinder	.4368451	0.71	.0609603	0.10
displace	.0287842	1.69	.0156661	0.93
horsepow	**-.0554725**	**-1.66**	**-.2849999**	**-3.91**
accel	.431885	2.23	.3059297	1.61
year	.2809669	1.04	.3410433	1.30
price	.0006226	3.19	.000649	3.45
wt*hp	**--------------**		**.0000745**	**3.51**
_cons	**27.85781**	**1.24**	**42.22624**	**1.91**

Car Mileage: No Interaction Effect

```
. regress   mpg weight cylinder displace horsepow accel year  price

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  7,   142) =   68.90
       Model |  6247.30769     7  892.472528           Prob > F      =  0.0000
    Residual |  1839.46671   142  12.9539909           R-squared     =  0.7725
-------------+------------------------------           Adj R-squared =  0.7613
       Total |   8086.7744   149  54.2736537           Root MSE      =  3.5992

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0123731   .0018463    -6.70   0.000    -.0160229   -.0087234
    cylinder |   .4368451   .6167177     0.71   0.480    -.7822891    1.655979
    displace |   .0287842   .0170537     1.69   0.094    -.0049278    .0624962
    horsepow |  -.0554725   .0333758    -1.66   0.099    -.1214501    .0105051
       accel |    .431885    .193399     2.23   0.027     .0495716    .8141984
        year |   .2809669   .2709474     1.04   0.302    -.2546449    .8165786
       price |   .0006226    .000195     3.19   0.002     .0002371    .0010081
       _cons |   27.85781    22.5288     1.24   0.218    -16.67736    72.39298
------------------------------------------------------------------------------
```

Car Mileage: With Interaction Effect


```
. regress   mpg weight cylinder displace horsepow accel year  price  wthp

      Source |       SS       df       MS              Number of obs =     150
-------------+------------------------------           F(  8,   141) =   66.61
       Model |  6394.74812     8  799.343515           Prob > F      =  0.0000
    Residual |  1692.02628   141  12.0001864           R-squared     =  0.7908
-------------+------------------------------           Adj R-squared =  0.7789
       Total |   8086.7744   149  54.2736537           Root MSE      =  3.4641

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0167031   .0021642    -7.72   0.000    -.0209816   -.0124246
    cylinder |   .0609603    .603188     0.10   0.920    -1.131501    1.253422
    displace |   .0156661   .0168351     0.93   0.354    -.0176157     .048948
    horsepow |  -.2849999   .0729368    -3.91   0.000     -.429191   -.1408088
       accel |   .3059297   .1895796     1.61   0.109    -.0688561    .6807156
        year |   .3410433   .2613443     1.30   0.194    -.1756165    .8577031
       price |    .000649   .0001878     3.45   0.001     .0002776    .0010203
        wthp |   .0000745   .0000212     3.51   0.001     .0000325    .0001165
       _cons |   42.22624    22.0676     1.91   0.058    -1.399893    85.85238
------------------------------------------------------------------------------
```

$$\frac{\partial MPG}{\partial wgt} = \beta_1 + \beta_8 \bullet hp$$

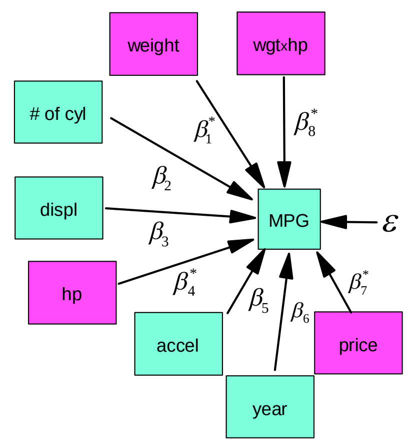
The marginal impact of a one pound change in a car's weight depends on the horsepower of the engine in the car.

$$\frac{\partial MPG}{\partial wgt} = \beta_1 + \beta_8 \bullet hp = -0.0167 + 0.0000745 \times hp$$

$$\frac{\partial MPG}{\partial hp} = \beta_4 + \beta_8 \bullet wgt = -0.285 + 0.0000745 \times wgt$$

	No Interaction		With Interaction	
	Coef.	t	Coef.	t
weight	**-.0123731**	**-6.70**	**-.0167031**	**-7.72**
cylinder	.4368451	0.71	.0609603	0.10
displace	.0287842	1.69	.0156661	0.93
horsepow	**-.0554725**	**-1.66**	**-.2849999**	**-3.91**
accel	.431885	2.23	.3059297	1.61
year	.2809669	1.04	.3410433	1.30
price	.0006226	3.19	.000649	3.45
wt*hp	**--------------**		**.0000745**	**3.51**
_cons	**27.85781**	**1.24**	**42.22624**	**1.91**

Handout on using metric variables in testing for interaction effects in multiple regression



The Question:
When comparing the total variance explained by the "base model" containing only the four significant variables to the variance explained by the "expanded model" which adds the insignificant variables, is the addition to explained variation large enough to say that this extra set of variables together has significantly raised total explained variance?

$H_0: \beta_2 = \beta_3 = \beta_5 = \beta_6 = 0$

H_a: at least one of these β not equal to zero

$$F^H_{n-K} = \frac{\left(RSS\{K-H\} - RSS\{K\}\right)/H}{\left(RSS\{K\}\right)/(n-K)} = \frac{(RSS_{restricted} - RSS_{unrestricted})/H}{RSS_{unrestricted}/n-K}$$

Residual Sum of Squares (RSS) is the sum of squared deviations of the residuals around their mean value of zero:

$$RSS\{K\} = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n}\left(Y_i - \hat{Y}\right)^2 = \sum_{i=1}^{n}\left(Y_i - \left[\hat{\beta}_0 + \sum_{k=1}^{K-1}\hat{\beta}_k X_{ik}\right]\right)^2$$

Remember, it's RSS that least squares regression seeks to minimize.

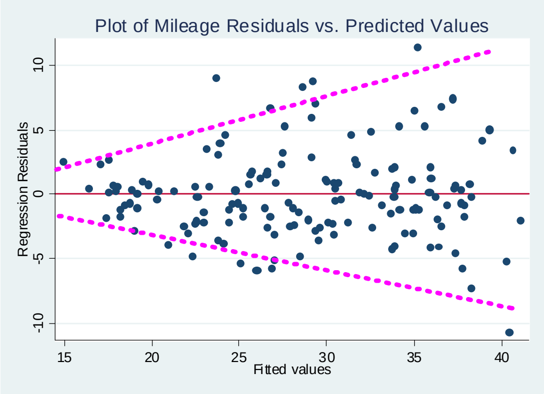
```
quietly regress mpg weight cylinder displace horsepow accel year
price  wthp
. test  cylinder= displace= accel= year=0


 ( 1)  cylinder - displace = 0
 ( 2)  cylinder - accel = 0
 ( 3)  cylinder - year = 0
 ( 4)  cylinder = 0

       F(  4,   141) =    1.50
            Prob > F =    0.2049
```

$H_0: \beta_4 = \beta_8 = 0$

H_a: at least one of these β not equal to zero

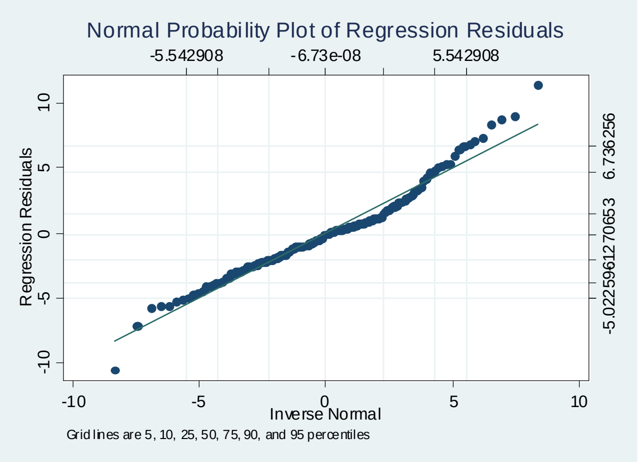
```
test  horsepow= wthp=0

 ( 1)  horsepow - wthp = 0
 ( 2)  horsepow = 0

       F(  2,   141) =    7.63
            Prob > F =    0.0007
```

```
predict mpghat

generate resid = mpg -mpghat


graph twoway scatter resid mpghat, yline(0)
ti("Plot of Mileage Residuals vs. Predicted
Values")
```

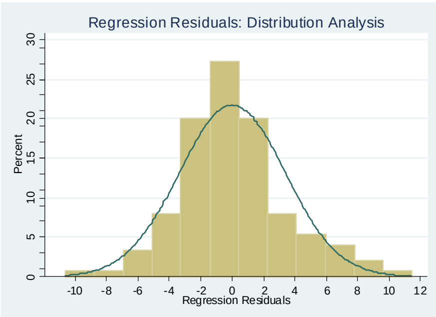


Automobile Mileage Data: Residual Plots

```
qnorm resid, grid ti(Normal Probability Plot of Regression
Residuals)
```

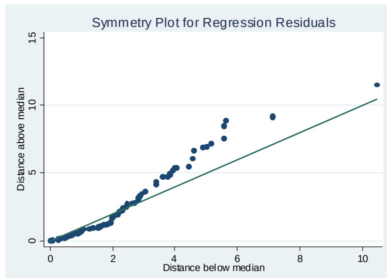


Residual Plots for mileage data regression



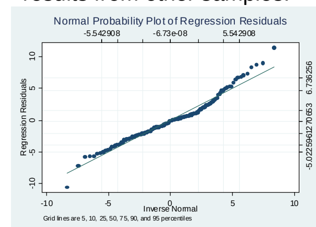

```
histogram resid, bin(12)
percent norm ti("Regression
Residuals: Distribution
Analysis") xlabel(-10 (2) 12)
ylabel(0 (5) 30) ytick(1 (2)
29)
```

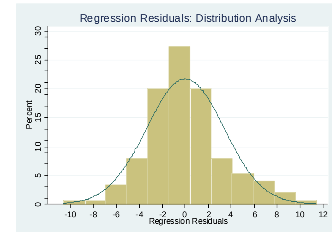
```
symplot resid, ti(Symmetry
Plot for Regression
Residuals)
```

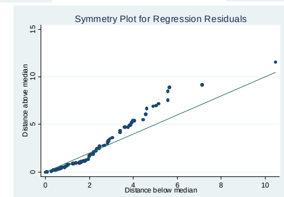


Nonnormality of the regression error term:

- We lose the justification for applying the *t* and the *F* distributions, especially with small samples.
- Since Ordinary Least Squares tends to be sensitive to outliers, heavy-tailed distributions can cause great sample-to-sample variation. I.e., our results from one random sample out of a population might not look very much like results from other samples.

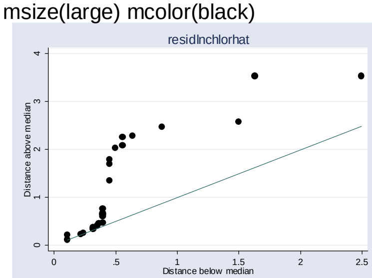




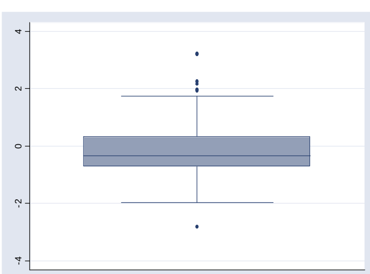



```
quietly regress lnchlor deep lndroad
deeproad
predict lnchlorhat
generate residlnchlorhat=lnchlor-
lnchlorhat
```

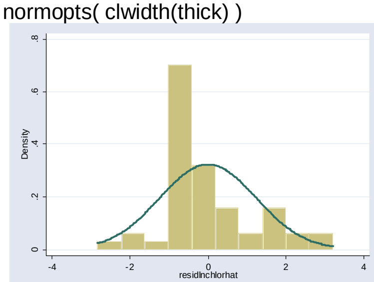
symplot residlnchlorhat, msymbol(circle) msize(large) mcolor(black)



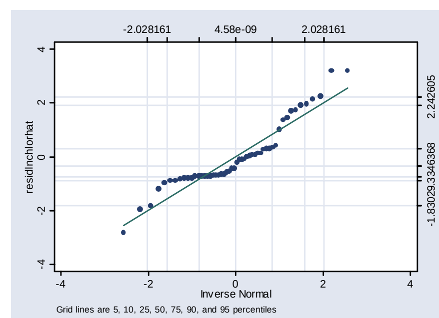
graph box residlnchlorhat, medtype(line)



histogram residlnchlorhat, bin(10) normal normopts(clwidth(thick))



qnorm residlnchlorhat, grid

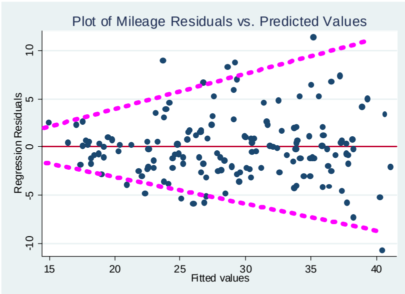


Some possible solutions to the nonnormality of residuals:

1. Transform the dependent variable.

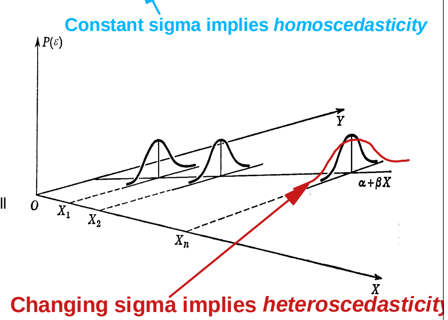
2. Transform one (or more) independent variables.

Possible transformations include: log transformations or power transformations.



$$E\left[\varepsilon_i \mid X\right] = E\left[\varepsilon_i\right] = 0 \text{ for all } i$$

$$E\left[\varepsilon_i \varepsilon_j\right] = \begin{cases} 0 & \text{for } i \neq j;\ i, j = 1, 2, \ldots, n \\ \sigma_\varepsilon^2 & \text{for } i = j;\ i, j = 1, 2, \ldots, n \end{cases}$$



Why is this a problem?

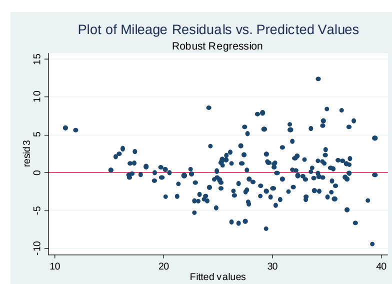
- Our initial assumption is that the variance of epsilon, conditional on the explanatory variables, is the same for all combinations is independent of the values of any of the independent variables. If this is not true, then the model exhibits **heteroscedasticity**.
- If heteroscedasticity is present, the coefficient estimates are unbiased, but the *estimated standard errors of the estimated coefficients are biased.*
- This means that hypothesis tests on regression coefficients (t-tests) and regressions (F-tests) will be incorrect, leading to incorrect inferences about our estimates.

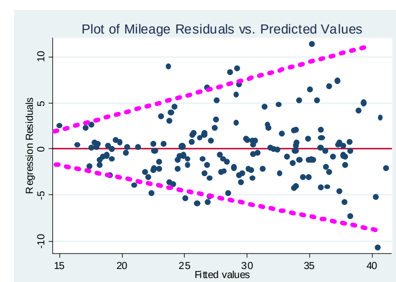
Are there ways to deal with heteroscedasticity?

- Stata can provide "heteroscedasticity robust" statistics after OLS estimation (see Hamilton "Robust Regression") , or
- We can uses Stata's "weighted least squares" to produce asymptotically unbiased estimates of standard errors. (See Hamilton)

```
. quietly regress  mpg cylinder displace horsepow accel year weight price

. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of mpg

         chi2(1)      =     4.32
         Prob > chi2  =   0.0377

. rreg mpg cylinder displace horsepow accel year weight price

   Huber iteration 1:  maximum difference in weights = .65098398
   Huber iteration 2:  maximum difference in weights = .11427817
   Huber iteration 3:  maximum difference in weights = .04099405
Biweight iteration 4:  maximum difference in weights = .29345885
Biweight iteration 5:  maximum difference in weights = .03798754
Biweight iteration 6:  maximum difference in weights = .00815723

Robust regression                                      Number of obs     =        150
                                                       F(  7,       142) =      69.74
                                                       Prob > F          =     0.0000

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    cylinder |   .1562774   .6051004     0.26   0.797    -1.039892    1.352447
    displace |     .03902   .0167325     2.33   0.021      .005943    .0720969
    horsepow |  -.0590311   .0327471    -1.80   0.074    -.1237659    .0057036
       accel |   .3803014   .1897559     2.00   0.047     .0051897     .755413
        year |   .3283337   .2658435     1.24   0.219    -.1971886     .853856
      weight |  -.0126884   .0018115    -7.00   0.000    -.0162694   -.0091074
       price |   .0006392   .0001913     3.34   0.001      .000261    .0010175
       _cons |   25.55935   22.10442     1.16   0.249     -18.1369     69.2556
------------------------------------------------------------------------------
```

```
. predict mpghat3
(option xb assumed; fitted values)

. generate resid3=mpg-mpghat3

. graph twoway scatter resid3 mpghat3, yline(0) ti("Plot of Mileage Residuals vs
> . Predicted Values") subtitle("Robust Regression")
```





Robust Regression

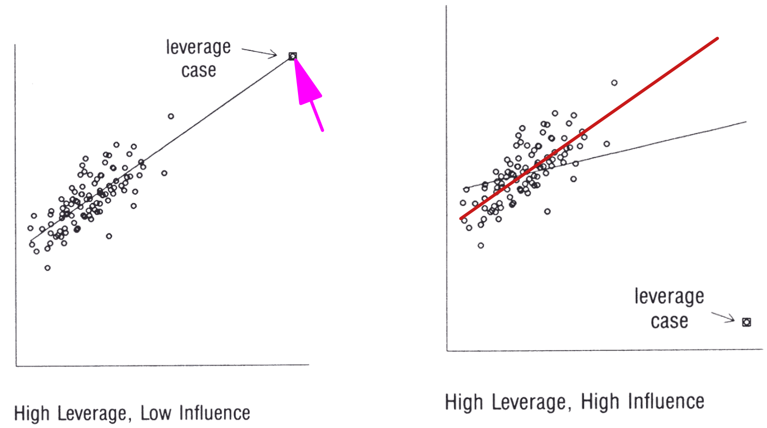
mpg	Coef.	Std. Err.	t	P>|t|
cylinder	.1562774	.6051004	0.26	0.797
displace	.03902	.0167325	2.33	0.021
horsepow	-.0590311	.0327471	-1.80	0.074
accel	.3803014	.1897559	2.00	0.047
year	.3283337	.2658435	1.24	0.219
weight	-.0126884	.0018115	-7.00	0.000
price	.0006392	.0001913	3.34	0.001
_cons	25.55935	22.10442	1.16	0.249

Displacement is now Significant

OLS Regression

mpg	Coef.	Std. Err.	t	P>|t|
cylinder	.4368451	.6167177	0.71	0.480
displace	.0287842	.0170537	1.69	0.094
horsepow	-.0554725	.0333758	-1.66	0.099
accel	.431885	.193399	2.23	0.027
year	.2809669	.2709474	1.04	0.302
weight	-.0123731	.0018463	-6.70	0.000
price	.0006226	.000195	3.19	0.002
_cons	27.85781	22.5288	1.24	0.218

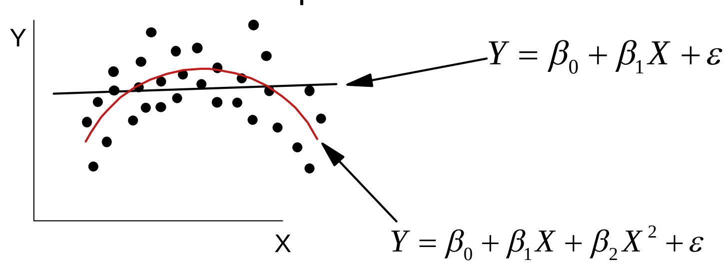
Influential Observations



Omitted Variable Bias -- Specification Error

The solution: Include all relevant explanatory variables. For this you need a strong theory of the causal process that you are trying to explain, or, lacking the appropriate variables, you need to run a well-controlled experiment.

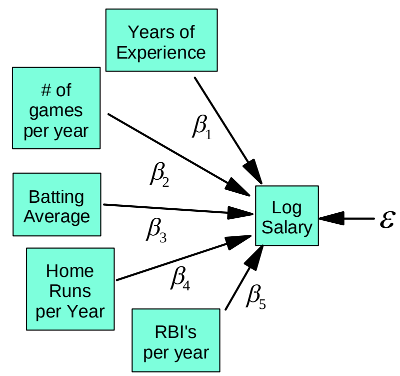
Functional Form -- Misspecification



Career Statistics for 353 Major League Baseball Players: 1993 Season

Goal: Predict determinants of their 1993 Salaries

The Model:



I'm using the logarithm of salary because of the wide variation in player salaries (Range: $109,000 to $6,329,213)

```
. regress lsalary  years gamesyr bavg hrunsyr rbisyr

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  5,   347) =  117.06
       Model |  308.989208     5  61.7978416           Prob > F      =  0.0000
    Residual |  183.186327   347  .527914487           R-squared     =  0.6278
-------------+------------------------------           Adj R-squared =  0.6224
       Total |  492.175535   352  1.39822595           Root MSE      =  .72658

------------------------------------------------------------------------------
     lsalary |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       years |   .0688626   .0121145     5.68   0.000     .0450355    .0926898
     gamesyr |   .0125521   .0026468     4.74   0.000     .0073464    .0177578
        bavg |   .0009786   .0011035     0.89   0.376    -.0011918     .003149
     hrunsyr |   .0144295    .016057     0.90   0.369    -.0171518    .0460107
      rbisyr |   .0107657    .007175     1.50   0.134    -.0033462    .0248776
       _cons |   11.19242   .2888229    38.75   0.000     10.62435    11.76048
------------------------------------------------------------------------------
. test  bavg= hrunsyr= rbisyr=0

 ( 1)  bavg - hrunsyr = 0
 ( 2)  bavg - rbisyr = 0
 ( 3)  bavg = 0

       F(  3,   347) =    9.55
            Prob > F =    0.0000
```

As a group the three variables are significant!

```
. correlate  years gamesyr bavg hrunsyr rbisyr
             |   years  gamesyr     bavg  hrunsyr   rbisyr
-------------+---------------------------------------------
       years |   1.0000
     gamesyr |   0.5624   1.0000
        bavg |   0.1973   0.3191   1.0000
     hrunsyr |   0.3802   0.6138   0.1906   1.0000
      rbisyr |   0.4871   0.8487   0.3291   0.8907   1.0000
```

Home Runs and RBI's are highly correlated!

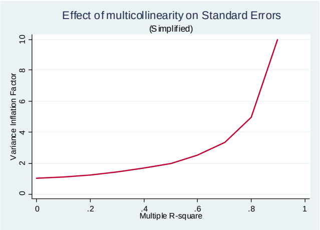
Multicollinearity refers to a situation in which there is a strong linear relationship among two or more independent variables in a multiple regression. In the more extreme cases, this linear correlation is so strong that the contribution of individual variables to the regression cannot be adequately ascertained.

$$\operatorname{var}\left(\hat{\beta}_j\right) = \frac{s^2}{(n-1)\operatorname{var}\left(X_j\right)} \bullet \frac{1}{1-R_j^2} \qquad \text{where}$$

- s^2 is the variance of the error term in the regression.
- *var(X_j)* is the variance of the independent variable *j*.
- R_j^2 is the multiple R^2 for the regression of X_j on the other covariates

$$\frac{1}{1-R_j^2} \text{ is the Variance Inflation Factor (VIF)}$$

As R_j^2 rises from 0 to one, the VIF approaches infinity.




```
. regress rbisyr hrunsyr gamesyr years bavg

      Source |       SS       df       MS              Number of obs =     353
-------------+------------------------------           F(  4,   348) = 1469.33
       Model |   173190.97     4  43297.7426           Prob > F      =  0.0000
    Residual |  10254.7377   348   29.467637           R-squared     =  0.9441
-------------+------------------------------           Adj R-squared =  0.9435
       Total |  183445.708   352   521.15258           Root MSE      =  5.4284

------------------------------------------------------------------------------
      rbisyr |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     hrunsyr |   1.998371   .0540007    37.01   0.000     1.892162     2.10458
     gamesyr |   .2979218   .0116611    25.55   0.000     .2749867    .3208569
       years |  -.1049775   .0903352    -1.16   0.246    -.2826491    .0726941
        bavg |   .0408626   .0079482     5.14   0.000       .02523    .0564953
       _cons |   -15.9307   1.981683    -8.04   0.000    -19.82828   -12.03312
------------------------------------------------------------------------------
```

The Variance Inflation Factor (VIF) for RBI's per year is: $\frac{1}{1-R_j^2} = \frac{1}{1-0.9441} = 17.889$

and: $\operatorname{var}\left(\hat{\beta}_j\right) = \frac{s^2}{(n-1)\operatorname{var}\left(X_j\right)} \bullet \frac{1}{1-R_j^2} = 0.000051 = Z \bullet 17.889$

This, in effect, means that RBI's per year and home runs per year are measuring much the same thing and Stata can't allocate explanatory power between them.

Dealing with Multicollinearity

- Use Stata's post estimation command *estat vif* to see if any coefficients show signs of trouble. E.g., for our baseball salary model we have:

Detection Rules of Thumb:

1. VIF for any variable is Greater Than or Equal to 10

2. Mean VIF for the entire regression is significantly greater than 1

```
. quietly regress lsalary rbisyr
hrunsyr gamesyr years bavg

. estat vif

    Variable |       VIF       1/VIF
-------------+----------------------
      rbisyr |     17.89    0.055901
     hrunsyr |      7.94    0.125913
     gamesyr |      6.10    0.163982
       years |      1.47    0.678753
        bavg |      1.20    0.834258
-------------+----------------------
    Mean VIF |      6.92
```

- Experiment with adding and deleting the suspect variables. Do standard errors or coefficient estimates change substantially?
- Employ one of the following coping strategies:

✓ Keep the variables in the equation, but understand that we cannot generalize (beyond the sample) about their separate effects.

✓ Drop one or more of the offending variables, since their information is mostly redundant.

✓ Combine the variables, since there is evidence that they are measuring the same thing.

✓ Collect more data. Multicollinearity is basically a problem of not enough information. Adding cases generally makes the coefficient estimates more precise, by (1) increasing n in the denominator of the formula, and (2) by increasing the variance of the independent variables.

$$\mathrm{var}\left(\hat{\beta}_j\right)=\frac{s^2}{(n-1)\,\mathrm{var}\left(X_j\right)}\bullet\frac{1}{1-R_j^2}$$

Larger *n* increases denominator

var(X_j) is likely to rise also